

BAYESIAN NETWORKS IN MULTIMODAL SPEECH RECOGNITION AND SPEAKER IDENTIFICATION

Ara V. Nefian and Lu Hong Liang

Intel Corporation
{ara.nefian, lu.hong.liang}@intel.com

ABSTRACT

Bayesian networks are statistical models that extend the framework of hidden Markov models (HMM) and allow for the analysis of multi modal signals such as audio-visual speech. Our recent results demonstrate the use of coupled HMM in audio-visual speech recognition and speaker identification. The increased performance of this model is due to its low complexity and its ability to describe both the audio-visual state asynchrony and natural dependency over time. The audio-visual speaker identification accuracy is enhanced in a late decision approach that integrates the audio-visual speech likelihood and the face likelihood computed using an embedded Bayesian network.

1. INTRODUCTION

Audio-visual systems, which include speech and speaker recognition, received a large interest in the last decade due to their increased accuracy and their robustness to acoustic and visual noise.

The system presented in this paper (Figure 1) uses a unified approach to audio-visual speech recognition (AVSR) and speaker identification (AVSI) that takes advantage of the correlation between the acoustic and visual speech. In our work, the sequence of visual features extracted from the mouth shape over time is combined with the features of acoustic speech to obtain a reliable recognition system. The motivation of our approach is supported by recent AVSR and AVSI systems that empirically demonstrate the strong correlation between acoustic and visual speech [6, 1, 22, 17]. The audio-visual fusion methods [22, 4] can be broadly grouped into two categories: feature fusion and decision systems. In feature fusion systems, such as the multi-stream HMM [18], or product HMM [21, 5], the observation vectors are obtained through the concatenation of acoustic and visual speech feature vectors. However, the audio and visual state synchrony assumed by these systems may not describe accurately the audio-visual speech generation [23, 3]. On the other hand, decision level fusion systems combine the recognition scores obtained independently for each modality. These systems often fail to entirely capture the

dependencies between the audio and video features [17]. The coupled hidden Markov model (CHMM) based AVSR

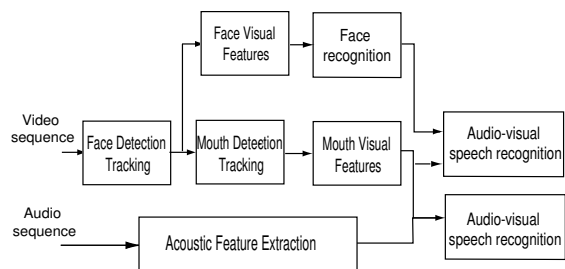


Figure 1: The overall system for audio visual speech recognition and speaker identification.

and AVSI systems described in this paper can be seen as an extension of the decision fusion system at phoneme-viseme level. A CHMM allows for audio-visual state-asynchrony as well as captures the natural conditional dependencies between the two modalities at the state level. Recently it has been shown that CHMMs outperform both the product and multi-stream HMM for the task isolated word audio-visual speech recognition [16].

2. THE AUDIO-VISUAL CHMM

A CHMM [2] can be seen as a collection of hidden Markov models (HMM), one for each data stream, where the hidden backbone nodes at time t for each HMM are conditioned by the backbone nodes at time $t - 1$ for all the related HMMs (Figure 2). Throughout this paper we will use CHMM with two channels one for audio and the other for visual observations. The parameters of a CHMM with two channels are defined below:

$$\begin{aligned}\pi_0^c(i) &= P(q_1^c = i) \\ b_t^c(i) &= P(\mathbf{O}_t^c | q_t^c = i) \\ a_{i|j,k}^c &= P(q_t^c = i | q_{t-1}^a = j, q_{t-1}^v = k)\end{aligned}$$

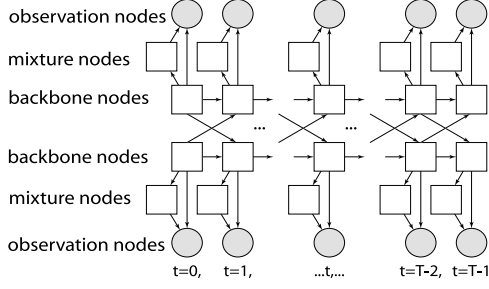


Figure 2: A time representation of the coupled HMM used in our AVSR and AVSI systems.

where $c \in \{a, v\}$ denotes the audio and visual channels respectively, and q_t^c is the state of the couple node in the c th channel at time t . In a continuous mixture with Gaussian components, the probabilities of the observed nodes are given by:

$$b_t^c(i) = \sum_{m=1}^{M_i^c} w_{i,m}^c N(\mathbf{O}_t^c, \mu_{i,m}^c, \mathbf{U}_{i,m}^c)$$

where \mathbf{O}_t^c is the observation vector at time t corresponding to channel c , and $\mu_{i,m}^c$ and $\mathbf{U}_{i,m}^c$ and $w_{i,m}^c$ are the mean and covariance matrix and mixture weight corresponding to the i th state, and m th mixture and the c th channel. M_i^c and are the number of mixtures corresponding to the i th state in the c th channel. In our AVSR system each CHMM describes one of the possible phoneme-viseme pairs as defined in [17], while in AVSI each CHMM describes one of the possible phoneme-viseme pairs for each person in the database.

3. TRAINING

The training of the CHMM parameters for AVSR starts with the Viterbi initialization and is followed by Expectation-Maximization training as described in [16]. The training of the CHMM parameters for the task of AVSI is performed in two stages. First, a speaker-independent background model (BM) is obtained for each CHMM corresponding to a viseme-phoneme pair. Next, the parameters of the CHMMs are adapted to a speaker specific model using a maximum a posteriori (MAP) method. To deal with the requirements of a continuous speech recognition systems, two additional CHMMs are trained to model the silence between consecutive words and sentences.

3.1. MAXIMUM LIKELIHOOD TRAINING OF THE BACKGROUND MODEL

In the first stage, the CHMMs for isolated phonem-viseme pairs are initialized using the Viterbi-based method described

in [13] followed by the estimation-maximization (EM) algorithm [10]. Each of the models obtained in the first stage are extended with one entry and one exit non-emitting states. The use of the non-emitting states also enforces the phoneme-viseme synchrony at the model boundaries (Figure 3). Next,

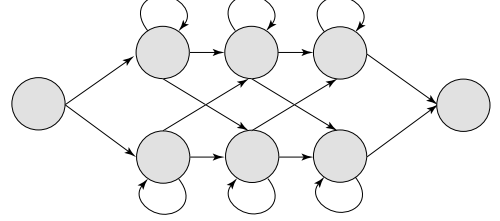


Figure 3: The state diagram of the coupled HMM used in our AVSR and AVSI systems.

the parameters of the CHMMs are refined through the embedded training of all CHMM from continuous audio-visual speech [10]. In this stage, the labels of the training sequences consist only on the sequence of phoneme-viseme with all boundary information being ignored. We will denote the mean, covariance matrices and mixture weights for mixture m , state i , and channel c of the trained CHMM corresponding to the background model as $(\mu_{i,m}^c)_{BM}$, $(\mathbf{U}_{i,m}^c)_{BM}$, and $(w_{i,m}^c)_{BM}$ respectively.

3.2. MAXIMUM A POSTERIORI ADAPTATION

In this stage of the training, the state parameters of the background model are adapted to the characteristics of each speaker in the database. The new state parameters for all CHMMs $\hat{\mu}_{i,m}^c$, $\hat{\mathbf{U}}_{i,m}^c$ and $\hat{w}_{i,m}^c$ are obtained through Bayesian adaptation [19]:

$$\hat{\mu}_{i,m}^c = \theta_{i,m}^c \mu_{i,m}^c + (1 - \theta_{i,m}^c) (\mu_{i,m}^c)_{BM} \quad (1)$$

$$\hat{\mathbf{U}}_{i,m}^c = \theta_{i,m}^c \mathbf{U}_{i,m}^c - (\mu_{i,m}^c)^2 + (\mu_{i,m}^c)_{BM}^2 + (1 - \theta_{i,m}^c) (\mathbf{U}_{i,m}^c)_{BM} \quad (2)$$

$$\hat{w}_{i,m}^c = \theta_{i,m}^c w_{i,m}^c + (1 - \theta_{i,m}^c) (w_{i,m}^c)_{BM} \quad (3)$$

where $\theta_{i,m}^c$ is a parameter that controls the MAP adaptation for mixture component m in channel c and state i . The sufficient statistics of the CHMM states corresponding to a specific user, $\mu_{i,m}^c$, $\mathbf{U}_{i,m}^c$ and $w_{i,m}^c$, are obtained using the EM algorithm from the available speaker dependent data as follows:

$$\mu_{i,m}^c = \frac{\sum_{r,t} \gamma_{r,t}^c(i, m) \mathbf{O}_{r,t}^c}{\sum_{r,t} \gamma_{r,t}^c(i, m)}$$

$$\mathbf{U}_{i,m}^c = \frac{\sum_{r,t} \gamma_{r,t}^c(i, m) (\mathbf{O}_{r,t}^c - \mu_{i,m}^c) (\mathbf{O}_{r,t}^c - \mu_{i,m}^c)^T}{\sum_{r,t} \gamma_{r,t}^c(i, m)}$$

$$w_{i,m}^c = \frac{\sum_{r,t} \gamma_{r,t}^c(i, m)}{\sum_{r,t} \sum_k \gamma_{r,t}^c(i, k)}$$

where

$$\begin{aligned} \gamma_{r,t}^c(i, m) &= \frac{\sum_j \frac{1}{P_r} \alpha_{r,t}(i, j) \beta_{r,t}(i, j)}{\sum_{i, j} \frac{1}{P_r} \alpha_{r,t}(i, j) \beta_{r,t}(i, j)} \times \\ &\times \frac{w_{i,m}^c N(\mathbf{O}_{r,t}^c | \mu_{i,m}^c, \mathbf{U}_{i,m}^c)}{\sum_k w_{i,k}^c N(\mathbf{O}_{r,t}^c | \mu_{i,k}^c, \mathbf{U}_{i,k}^c)} \end{aligned}$$

and $\alpha_{r,t}(i, j) = P(\mathbf{O}_{r,1}, \dots, \mathbf{O}_{r,t} | q_{r,t}^a = i, q_{r,t}^v = j)$ and $\beta_{r,t}(i, j) = P(\mathbf{O}_{r,t+1}, \dots, \mathbf{O}_{r,T_r} | q_{r,t}^a = i, q_{r,t}^v = j)$ are the forward and backward variables respectively [10] computed for the r th observation sequences

$\mathbf{O}_{r,t} = [(\mathbf{O}_{r,t}^a)^T, (\mathbf{O}_{r,t}^v)^T]^T$. The adaptation coefficient is obtained as

$$\theta_{i,m}^c = \frac{\sum_{r,t} \gamma_{r,t}^c(i, m)}{\sum_{r,t} \gamma_{r,t}^c(i, m) + \delta}$$

where δ is the relevance factor which is set $\delta = 16$ in our experiments. Note that as more speaker dependent data for a mixture m of state i and channel c becomes available, the contribution of the speaker specific statistics to the MAP state parameters increases (Equations 1- 3). On the other side, when less speaker specific data is available, the MAP parameters are very close to the parameters of the background model.

4. THE FACE MODEL

While HMM are very successful in speech or gesture recognition, an equivalent two-dimensional HMM for images has been shown to be impractical due to its complexity [7]. Figure 4a shows a graph representation of the 2D HMM with the square nodes representing the discrete hidden nodes and the circles describing the continuous observation nodes. In recent years, several approaches to approximate a 2D HMM with computationally practical models have been investigated [20, 14, 8]. In this paper, the face images are modeled using an embedded HMM (EHMM) [15]. The EHMM used for face recognition is a hierarchical statistical model with two layers of discrete hidden nodes (one layer for each data dimension) and a layer of observation nodes. In an EHMM both the ‘‘parent’’ and ‘‘child’’ layer of hidden nodes are described by a set of HMMs (Figure 4b). The states of the HMM in the ‘‘parent’’ and ‘‘child’’ layers are referred to as the *super states* and the states of the model respectively. The hierarchical structure of the EHMM or the embedded Bayesian networks [14] in general reduces significantly the complexity of these models compared to the 2D HMM.

The sequence of observation vectors for an EHMM are obtained from a window that scans the image from left to right and top to bottom as shown in Figure 4c. Using the images in the training set, an EHMM is trained for each person in the database by means of the EM algorithm described in [15]. Recognition is carried out via the Viterbi decoding algorithm [12].

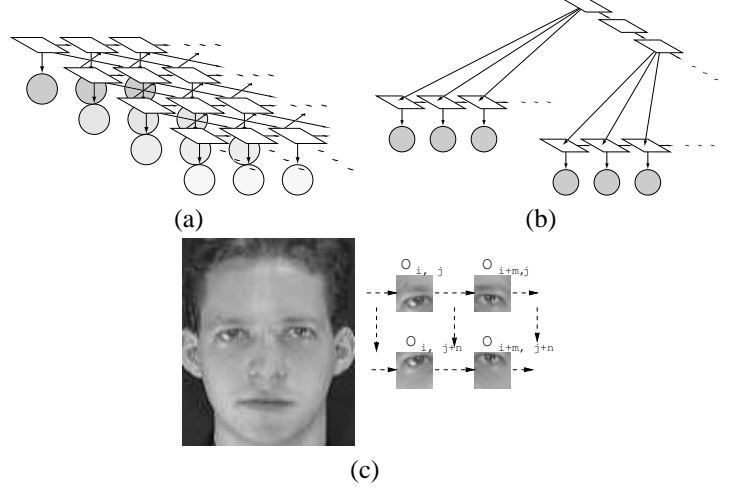


Figure 4: A 2D HMM for face recognition (a), an embedded HMM (b) and the facial feature block extraction for face recognition (c).

5. RECOGNITION

The AVSR is performed via a Viterbi algorithm described in [10]. The task of AVSI is performed in two stages. First the face likelihood and the audio-visual speech likelihood are computed separately. To deal with the variation in the relative reliability of the audio and visual features of speech at different levels of acoustic noise, we modified the observation probabilities used in decoding such that $\tilde{b}_k^c(i) = [b_k^c]^\lambda$, $c \in \{a, v\}$ where the audio and video stream exponents λ_a and λ_v satisfy $\lambda_a, \lambda_v \geq 0$ and $\lambda_a + \lambda_v = 1$. Then the overall matching score of the audio-visual speech and face model is computed as

$$L(\mathbf{O}^f, \mathbf{O}^a, \mathbf{O}^v | k) = \lambda_f L(\mathbf{O}^f | k) + \lambda_{av} L(\mathbf{O}^a, \mathbf{O}^v | k) \quad (4)$$

where $\mathbf{O}^a, \mathbf{O}^v$ and \mathbf{O}^f are the acoustic speech, visual speech and facial sequence of observations, $L(*|k)$ denotes the observation likelihood for the k th person in the database and $\lambda_f, \lambda_{av} \geq 0, \lambda_f + \lambda_{av} = 1$ are some weighting coefficients for the face and audio-visual speech likelihoods.

6. EXPERIMENTAL RESULTS

In our experiments the acoustic observation vectors consist of 13 MFCC coefficients, extracted from a window of 25.6 ms, with an overlap of 15.6 ms, with their first and second order time derivatives. The visual features are obtained from the mouth region through a cascade algorithm described in [9]. The extraction of the visual features starts with the neural network based face detection system followed by the detection and tracking of the mouth region using a set of support vector machine classifiers (Figure 5).

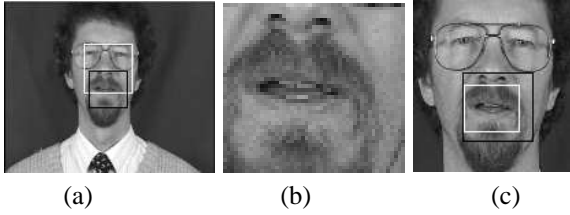


Figure 5: (a) An example of the face detection (white rectangle), and the estimated region of search for the mouth (black rectangle). (b) The estimated region of search for the mouth, enlarged. (c) The mouth detection result (white rectangle) from the initial region of search for the mouth (black rectangle).

The pixels in the mouth region are mapped to a 32-dimensional feature space using the principal component analysis. Then, blocks of 15 consecutive visual observation vectors are concatenated and projected on a 13 class, linear discriminant space. Finally, the resulting vectors of size 13, their first and second order time derivatives are used as the visual observation sequences. The audio and visual features are integrated using a CHMM with three states in both the audio and video chains with no back transitions (Figure 3). Each state has 32 mixture components using diagonal covariance matrix.

We tested the audio-visual continuous speech recognition system described here on the XM2VTS database [11]. We used a set of 1450 digit enumeration sequences captured from 200 speakers for training and a set of 700 sequences from other 95 speakers for decoding. The training sequences are recorded with "clean" audio. The audio data of the testing sequences is corrupted with several levels of white noise to allow the study of AVSR under less constrained acoustic conditions. Figure 6 compare the WER of our current AVSR system with an audio only speech recognition system.

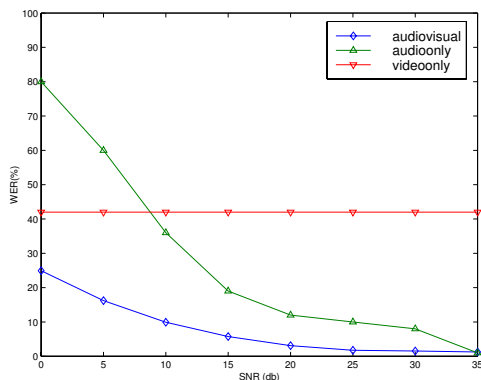


Figure 6: The word error rate of the audio-only, visual-only and audio-visual speech recognition system at different levels of SNR.

Our AVSI system uses the same acoustic and visual features as the AVSR system. The facial features are obtained using a sampling window of size 8×8 with 75% overlap between consecutive windows. The observation vectors corresponding to each position of the sampling window consist of a set of 2D discrete Cosine transform (2D DCT) coefficients. Specifically, we used nine 2D DCT coefficients obtained from a 3×3 region around the lowest frequency in the 2D DCT domain. The faces of all people in the database are modeled using EHMM with five super states and 3,6,6,6,3 states per super state respectively. Each state of the hidden nodes in the "child" layer of the EHMM is described by a mixture of three Gaussian density functions with diagonal covariance matrices.

The audio-visual speaker identification system presented in this paper was tested on digit enumeration sequences from the XM2VTS database [11]. For parameter adaptation we used four training sequences from each of the 87 speakers in our training set while for testing we used 320 sequences. To evaluate the behavior of our speaker identification system in environments affected by acoustic noise, we corrupted the testing sequences with white Gaussian noise at different SNR levels, while we trained on the original clean acoustic sequences. Figure 7 shows the error rate of the audio-only $(\lambda_v, \lambda_f) = (0.0, 0.0)$, video-only $(\lambda_v, \lambda_f) = (1.0, 0.0)$, audio-visual $(\lambda_f = 0.0)$ and face-audio-visual $(\lambda_v, \lambda_f) = (0.5, 0.3)$ speaker identification system at different SNR levels.

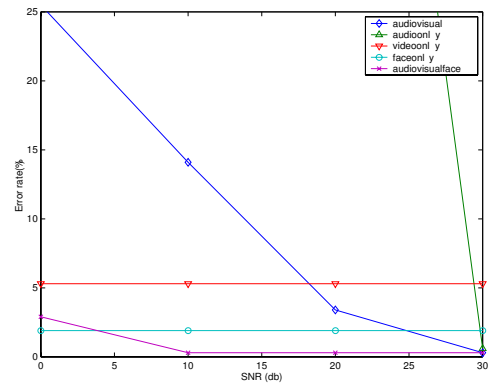


Figure 7: The error rate of the audio-only $(\lambda_v, \lambda_f) = (0.0, 0.0)$, video-only $(\lambda_v, \lambda_f) = (1.0, 0.0)$, audio-visual $(\lambda_f = 0.0)$ and face+audio-visual $(\lambda_v, \lambda_f) = (0.5, 0.3)$ speaker identification system.

7. CONCLUSIONS

In this paper we investigated the use of CHMM for the task of audio visual speech recognition and speaker identification. For both systems the visual features are obtained from

the mouth movement through the same cascade algorithm. In addition, the accuracy of the AVSI system is enhanced through a late decision fusion with a Bayesian network face recognition system. The use of strongly correlated acoustic and visual temporal features makes the current systems more robust to acoustic noise, and increases their accuracy by jointly modelling the multi-modal sequences.

Future research will study the performance of the current system for different types of acoustic noise as well as visual noise. The similarity between the visual features and fusion models used in the AVSR and the text dependent AVSI systems presented here allows for the integration of the two systems in a text independent AVSI system.

8. REFERENCES

- [1] A.Senior, C.Neti, and B.Maison. On the use of visual information for improving audio-based speaker recognition. In *In Proc. of Audio Visual Speech Processing Conference*, 1999.
- [2] M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 994–999, 1997.
- [3] T. Chen. Audiovisual speech processing. *Signal Processing Magazine*, 18:9–21, January 2001.
- [4] C. Chibelushi, F. Deravi, and S.D. Mason. A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, 4(1):23–37, March 2002.
- [5] S. Dupont and J. Luetttin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2:141–151, September 2000.
- [6] J.Luetttin, N.Thacker, and S.Beer. Speaker identification by lipreading. In *In Proc. of International Conference on Spoken Language Processing*, pages 62–64, 1996.
- [7] S. Kuo and O.E. Agazzi. Keyword spotting in poorly printed documents using pseudo 2-D Hidden Markov Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):842–848, August 1994.
- [8] Jia Lia, A. Najmi, and R.M. Gray. Image classification by a two-dimensional hidden markov model. *IEEE Transactions on Signal Processing*, 48(2):517–533, February 2000.
- [9] L. Liang, X. Liu, X. Pi, Y. Zhao, and A. V. Nefian. Speaker independent audio-visual continuous speech recognition. In *International Conference on Multimedia and Expo*, volume 2, pages 25–28, 2002.
- [10] X. Liu, L. Liang, Y. Zhao, X. Pi, and A. V. Nefian. Audio-visual continuous speech recognition using a coupled hidden Markov model. In *International Conference on Spoken Language Processing*, 2002.
- [11] J. Luetttin and G. Maitre. Evaluation protocol for the XM2FDB database. In *IDIAP-COM 98-05*, 1998.
- [12] A. V. Nefian and M. H. Hayes. Face recognition using an embedded HMM. In *Proceedings of the IEEE Conference on Audio and Video-based Biometric Person Authentication*, pages 19–24, March 1999.
- [13] A. V. Nefian, L. Liang, X. Pi, X. Liu, and C. Mao. An coupled hidden Markov model for audio-visual speech recognition. In *International Conference on Acoustics, Speech and Signal Processing*, 2002.
- [14] Ara V. Nefian. Embedded Bayesian networks for face recognition. In *IEEE International Conference on Multimedia and Expo*, volume 2, pages 25–28, 2002.
- [15] Ara V. Nefian and Monson H. Hayes III. Maximum likelihood training of the embedded HMM for face detection and recognition. In *IEEE International Conference on Image Processing*, volume 1, pages 33–36, 2000.
- [16] A.V. Nefian, L. Liang, X. Liu, X. Pi, C. Mao, and K. Murphy. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Applied Signal Processing, special issue on Audio Visual Signal Processing*, 2002, to appear.
- [17] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio visual speech recognition. In *Final Workshop 2000 Report*, 2000.
- [18] G. Potamianos, J. Luetttin, and C. Neti. Asynchronous stream modeling for large vocabulary audio-visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 169–172, 2001.
- [19] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using an adapted gaussian mixture model. *Digital Signal Processing*, 10:19–41, 2000.
- [20] F. Samaria. *Face Recognition Using Hidden Markov Models*. PhD thesis, University of Cambridge, 1994.
- [21] L.K. Saul and M.L. Jordan. Boltzmann chains and hidden Markov models. In G. Tesauro, David S. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. The MIT Press, 1995.
- [22] S.B.Yacouband, S.Luetttin, J.Jonsson, K.Matas, and J.Kittler. Audio-visual person verification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 580–585, 1999.
- [23] P. Teissier, J. Rober-Ribes, J.-L. Schwartz, and A. Guerin-Dugue. Comparing models for audio-visual fusion in a noisy vowel recognition task. *IEEE Transactions on Speech and Audio and Signal Processing*, 7:629–642, 1999.